Metagenome Assembly & Binning

also sometimes called *de novo* metagenome analysis

STAMPS 2022

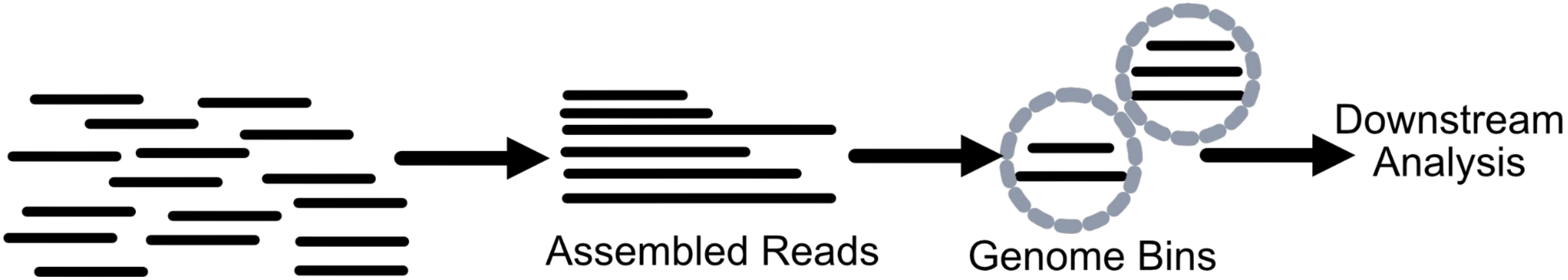
Taylor Reiter, PhD

Metagenome Assembly & Binning mostly for short reads

also sometimes called *de novo* metagenome analysis

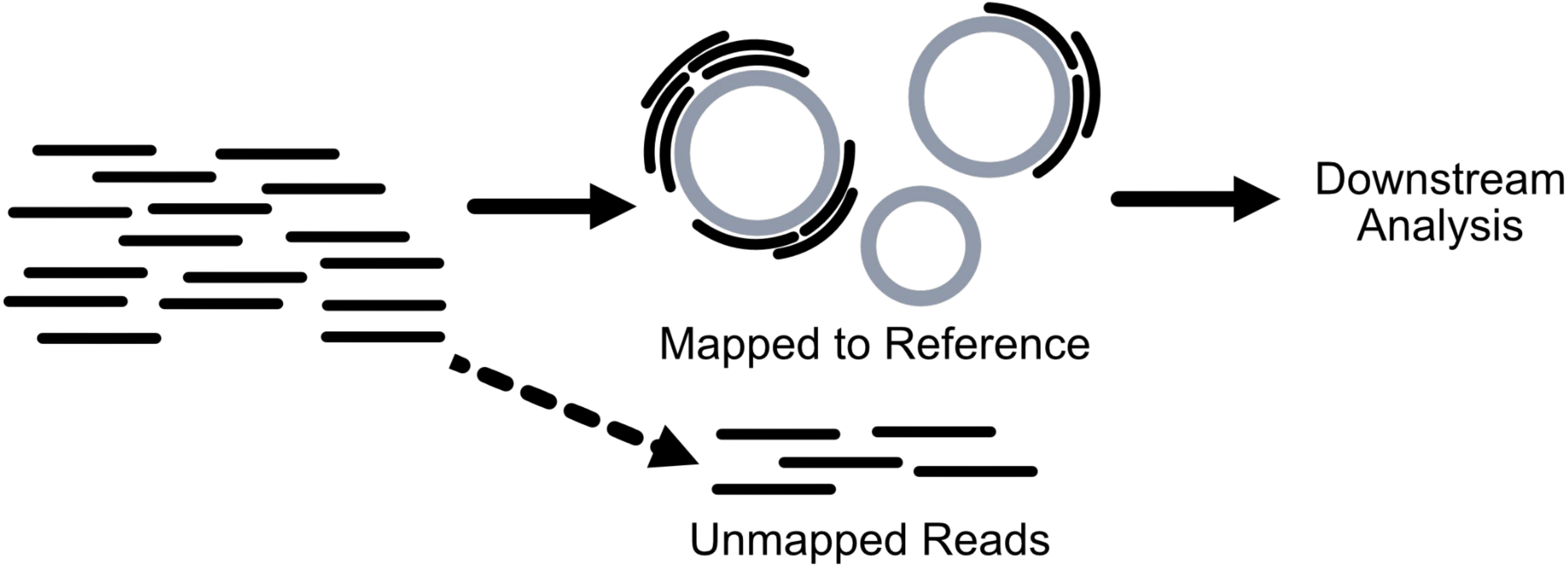
STAMPS 2022

Assembly & Binning

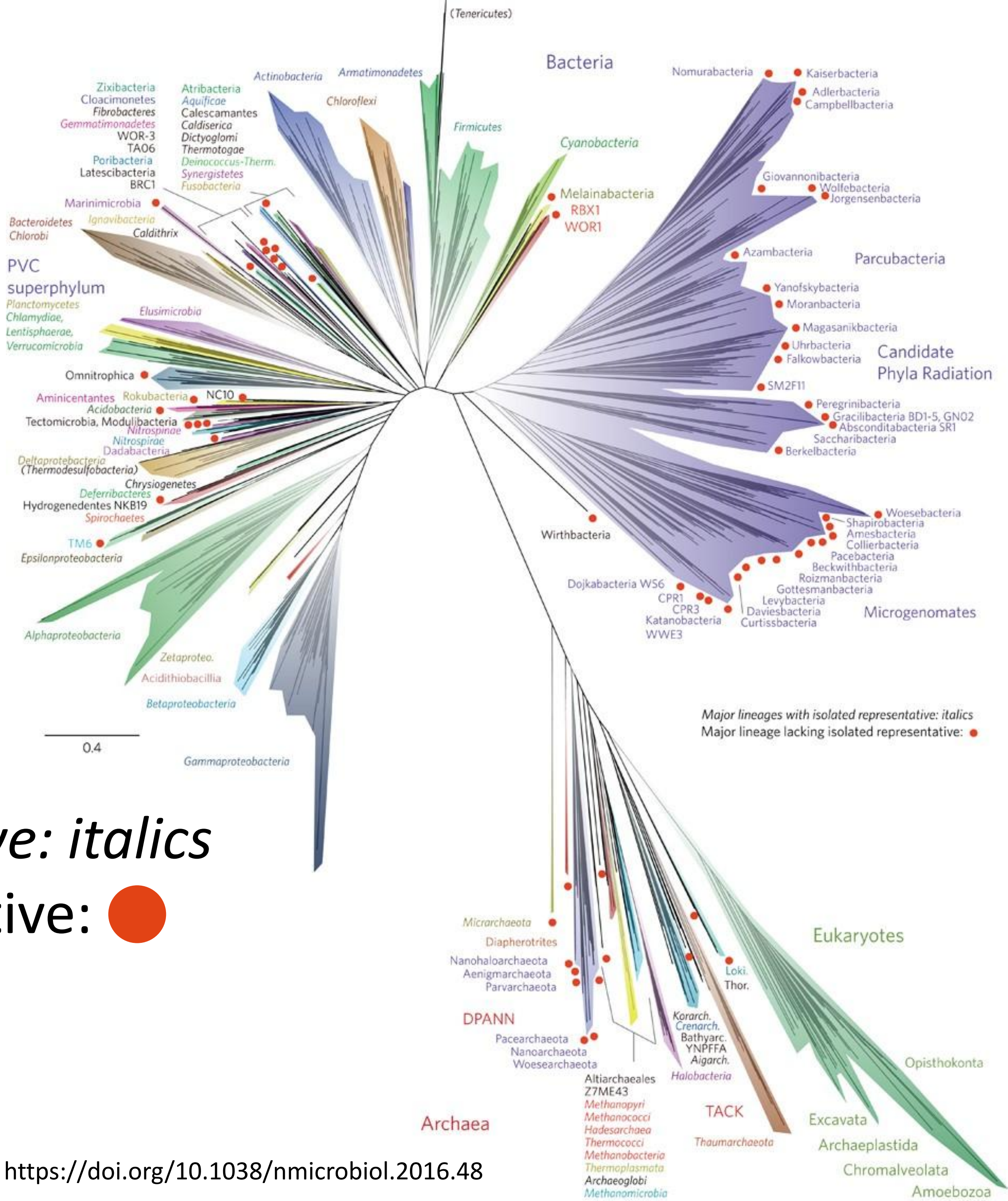


Why do we do *de novo* metagenome analysis?

Why do we do *de novo* metagenome analysis: reference databases are incomplete and we have to do something



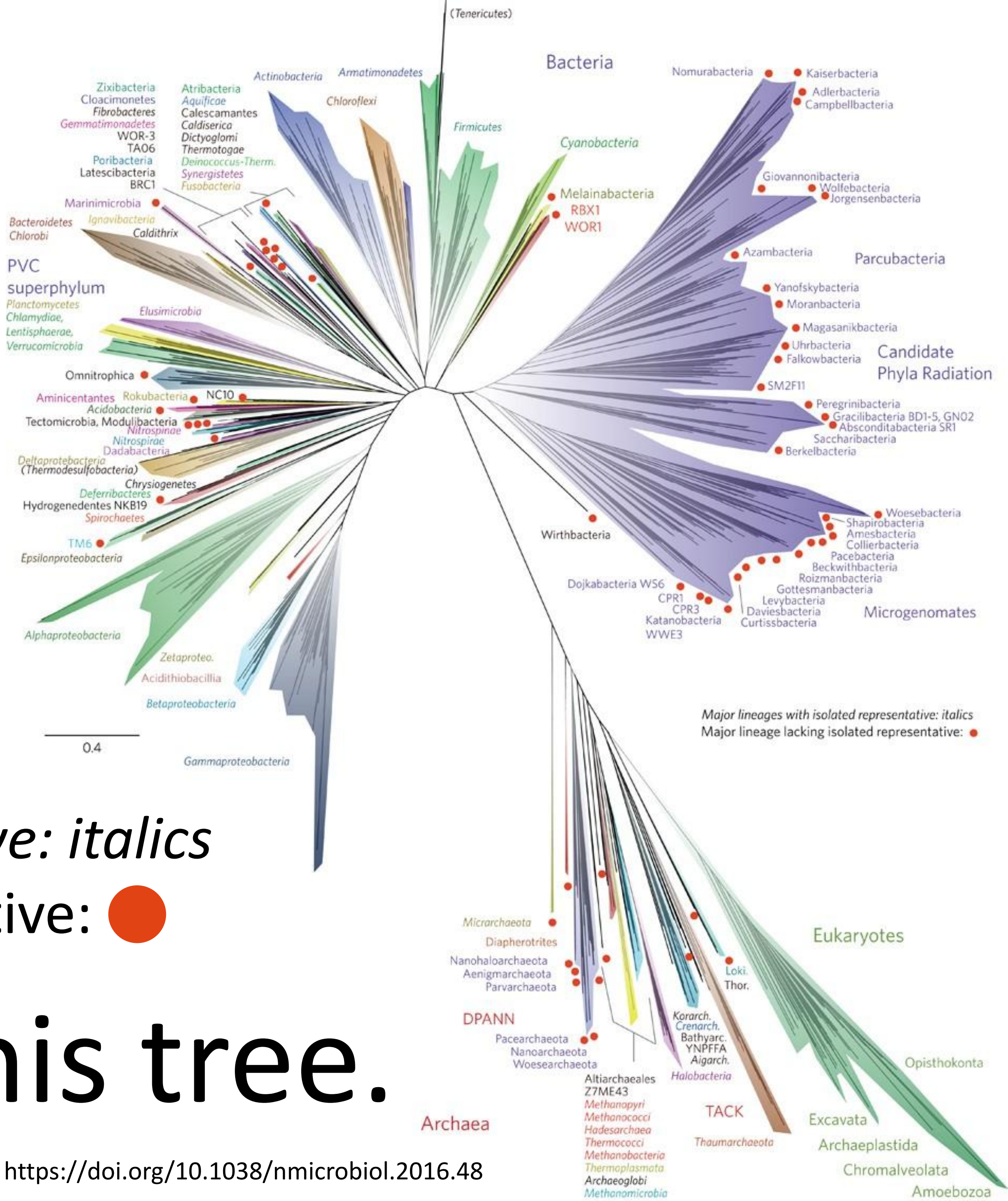
What has *de novo* metagenome analysis given us?



Major lineages with isolated representative: italics

Major lineage lacking isolated representative: ●

Hug, L., Baker, B., Anantharaman, K. *et al.* A new view of the tree of life. *Nat Microbiol* **1,** 16048 (2016). https://doi.org/10.1038/nmicrobiol.2016.48



Major lineages with isolated representative: italics

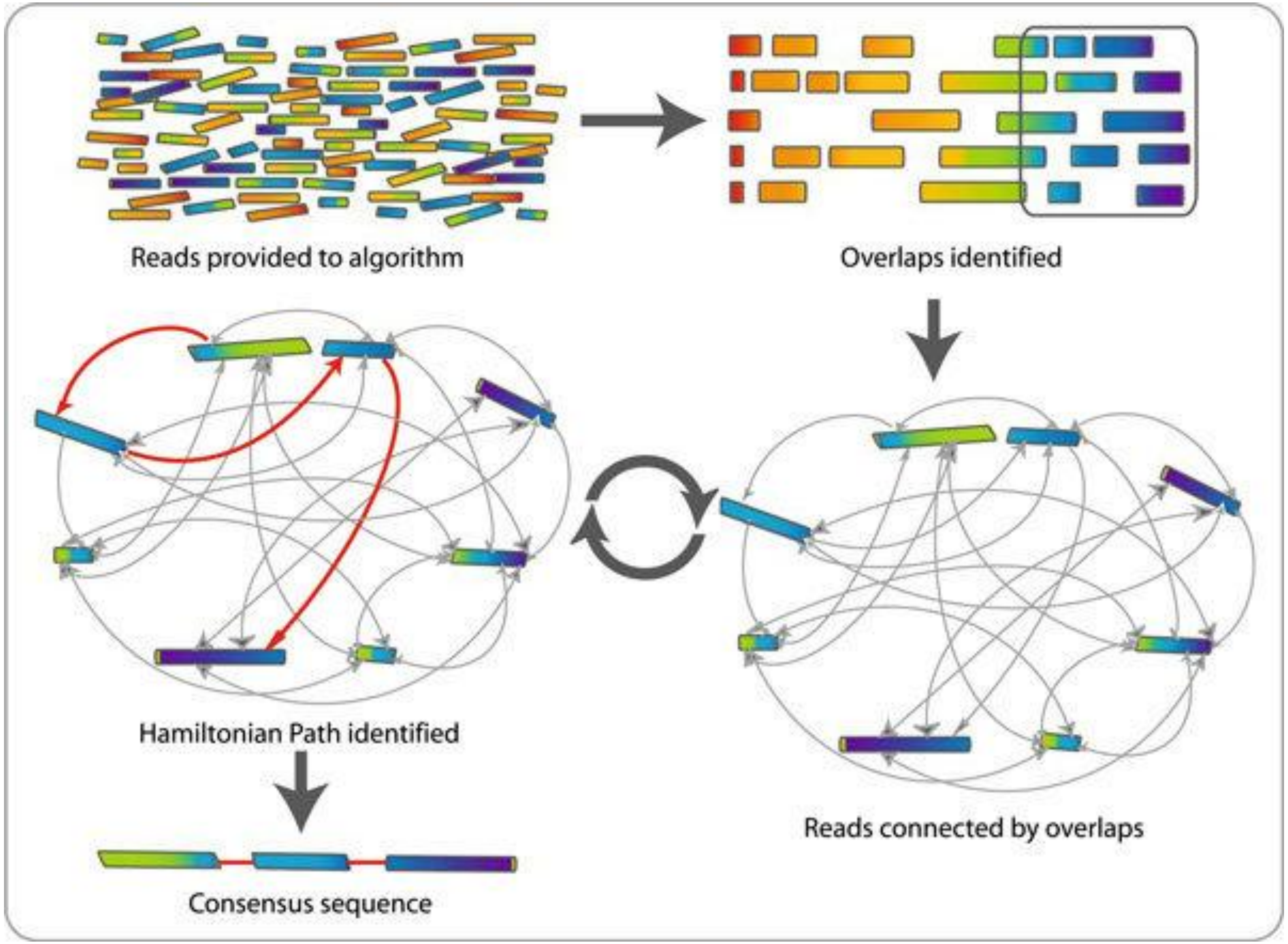
Major lineage lacking isolated representative: ●

There are **68** ● on this tree.

Hug, L., Baker, B., Anantharaman, K. *et al.* A new view of the tree of life. *Nat Microbiol* **1,** 16048 (2016). https://doi.org/10.1038/nmicrobiol.2016.48

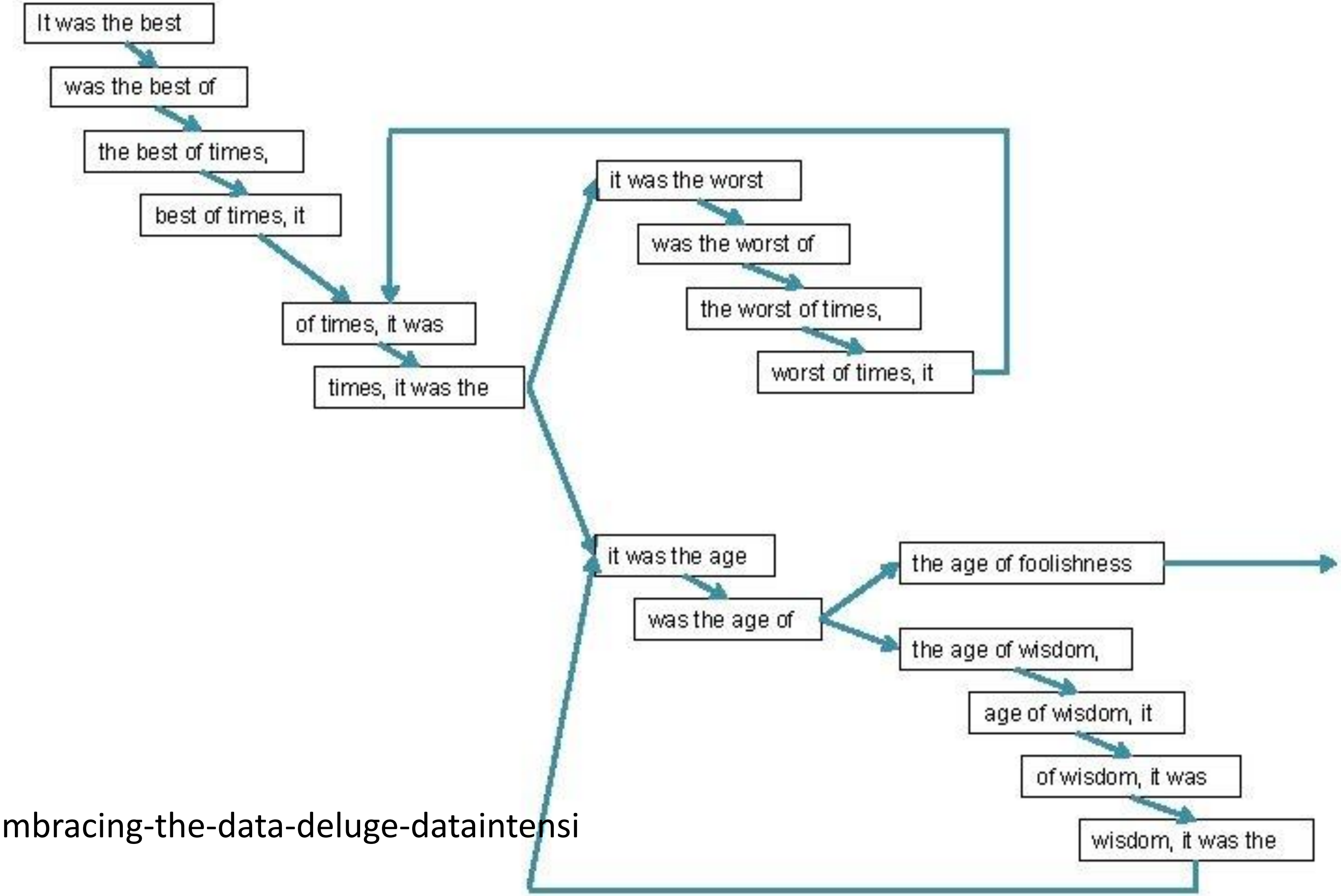
How does assembly work?

It was the best of times, it was the worst of times


Reads provided to algorithm
Overlaps identified
Reads connected by overlaps
Hamiltonian Path identified
Consensus sequence

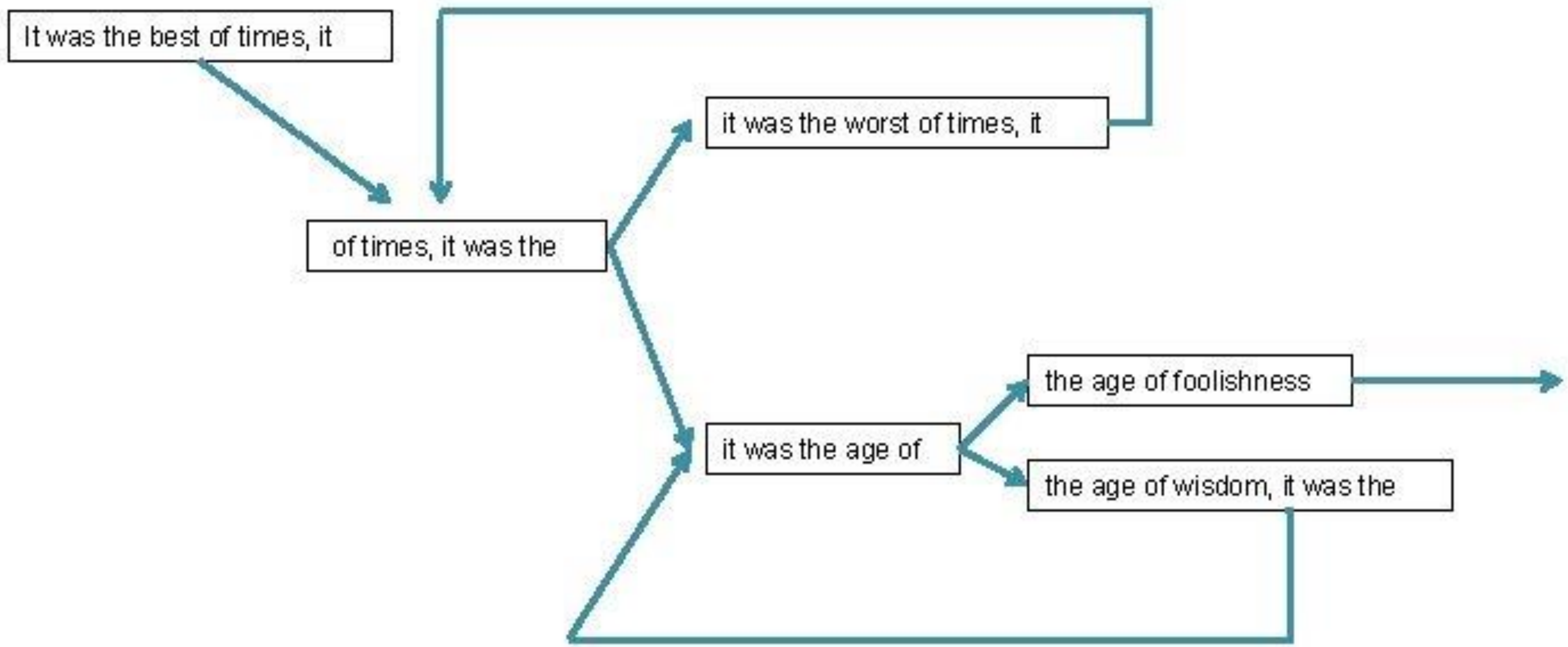
Commins, J., Toft, C. & Fares, M.A. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. *Biol Proced Online* **11,** 52 (2009). https://doi.org/10.1007/s12575-009-9004-1

Common assembly strategies: **de Bruijn graph methods**



https://slidetodoc.com/embracing-the-data-deluge-dataintensive-computing-for-the/

Common assembly strategies: de Bruijn graph methods



It was the best of times, it

of times, it was the

it was the worst of times, it

it was the age of

the age of foolishness

the age of wisdom, it was the

https://slidetodoc.com/embracing-the-data-deluge-dataintensive-computing-for-the/

Tools that do assembly not an exhaustive list

Overlap layout consensus

- ?
- Long read assemblers?

Greedy

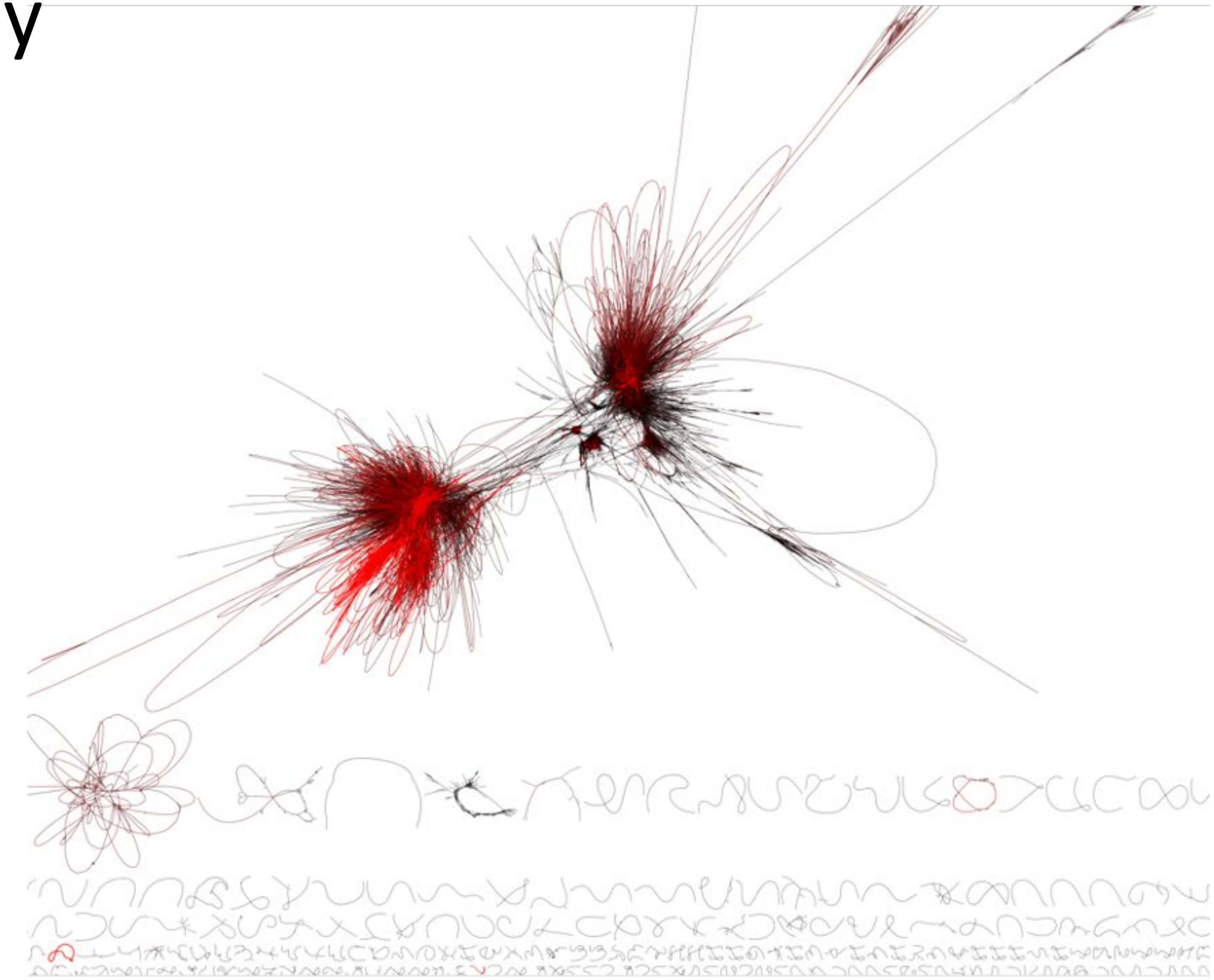
- PLASS

de Bruijn Graph

- (meta)SPAdes
- Megahit
- IDBA-UD
- MetaVelvet
- Ray Meta

Metagenome assembly graphs in the wild

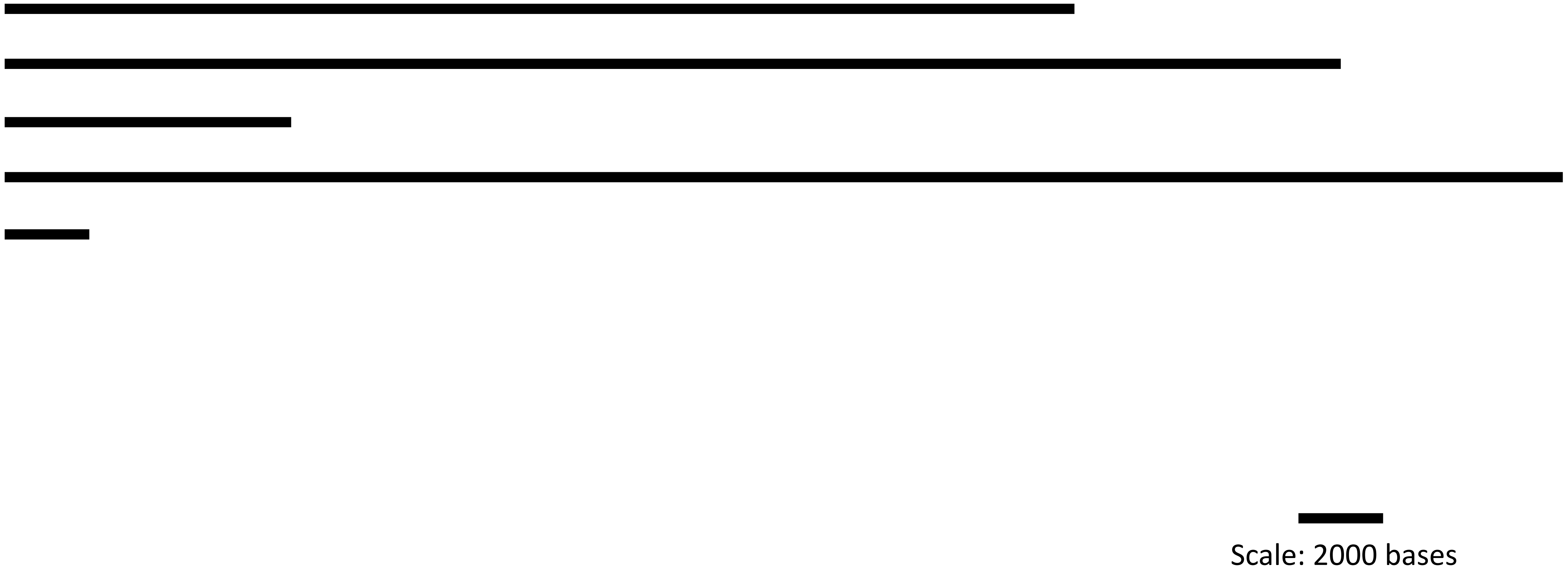
Mouse gut metagenome



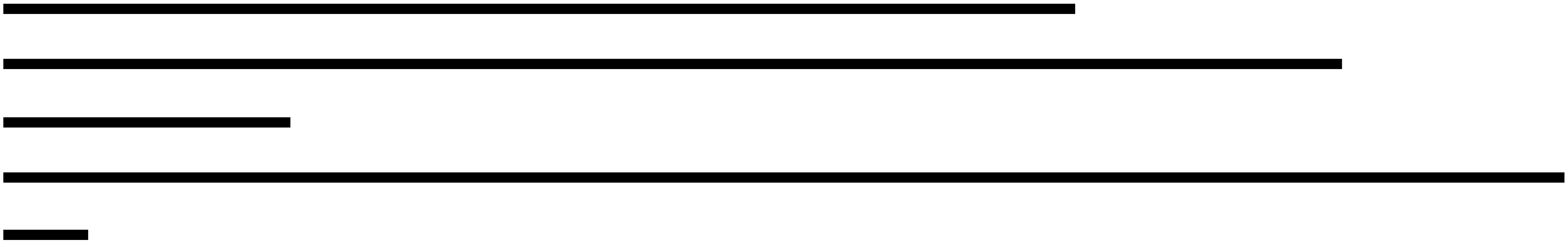
@SilasKieser https://twitter.com/SilasKieser/status/1308752555795779585/photo/1

What does an assembly look like?

Metagenome assembly



Metagenome assembly



```
>SRR8859675_0
TATTAATTGGAGCCGTAGCTTCAGAGAAAATACCAGCTACGGTCTTTTCATTTTTATTAGAATTATCGTC
TATACTAACGTCATTTAACTTTTCAGTAATTGCCTTCTCTACTTGTCTAATAAAATTAAATTTATCTTGA
TTTAAACCACTAATTTCCATTAAATTTTCTACTATATTTTTAACCTTTTCTGCGTCATCTGTATTTAAAA
GATTTTTAATATTTGATTTTAGTACAGTTTTCATTCCTCTAAGAGTTGTAATCGTATCTGCTTCATTGTA
AAATTTCTCTGAAAATTCTTCTATTTGCATTTTTCCTCCTAAGAAAACGTTAATTTGCCATTAATTGATA
TACATTCCCCATTATAATTGTATATTTTTTGGTTTGTACTTGCAAGTATGAATTCATTATCAGTCTTTAT
TGATTCTTGATTTAATAATGCATCATATTTGCCACATTTTAAATATTCGTATCCTTTTATATCTGCACAT
>SRR8859675_1
GCGGCAGGCATACCAGCTCGACATATCATAACGAGAGATACTTGTCCTCCAGGTCAGCAGTCAGCAGAAC
AATGGGTTCGGCAACGTTTTGAACAAACGCTGAAGACATTCAGAAGCAAGCATGGGCAGGGTCGGAAGAT
ATGCCTTATCGTGATGATCGATGCTGATCGTCATACCCCTGAAGAACGCAGGAAACAGTTACAGAAGAAT
ATAAAAAGAGAAAACGGAGAGCCGATTGGAATTTTTGTTCCAGCGAGAAACATCCAGAGCTGGATGGCCT
```

Scale: 2000 bases

When does assembly fail?

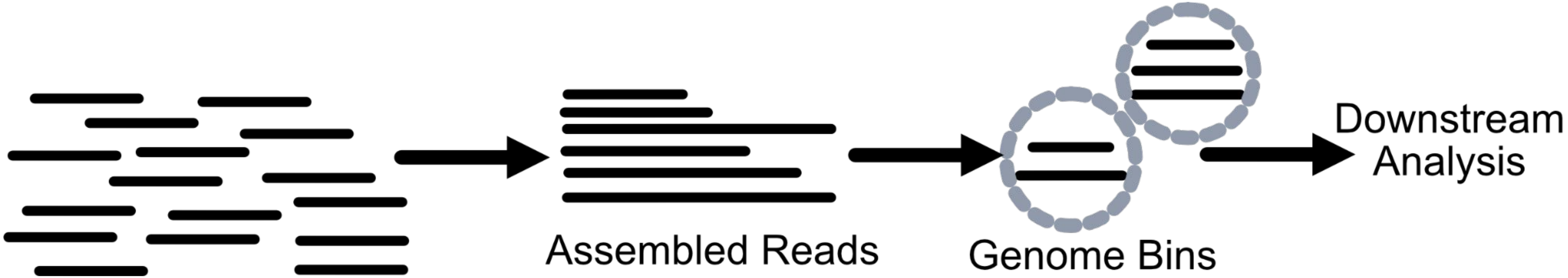
How do we evaluate an assembly?

Binning

We have an assembly. Now what? May we haz genomes?

Everything’s made up and the points don’t matter

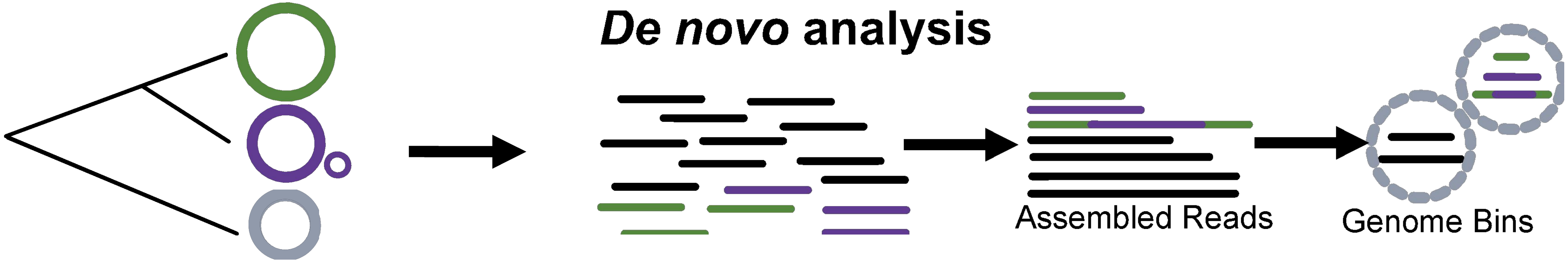
Assembly & Binning



How do we bin?

How do we evaluate binning?

Why do we evaluate binning?



Where does binning fail most frequently?

🥳 Tutorial 🥳

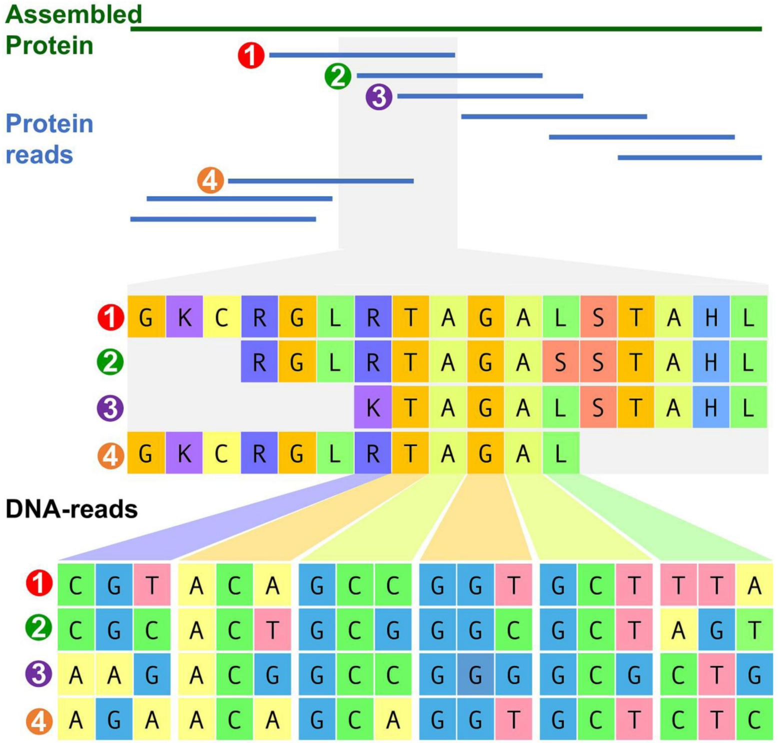
Strategies for when assembly fails

How many of reads did not assemble?

Assemble in protein space

Pros

- With less microdiversity in protein space -> more assembly



1 G K C R G L R T A G A L S T A H L
2 R G L R T A G A S S T A H L
3 K T A G A L S T A H L
4 G K C R G L R T A G A L

DNA-reads

1 C G T A C A G C C G G T G C T T T A
2 C G C A C T G C G G G C G C T A G T
3 A A G A C G G C C G G G G C G C T G
4 A G A A C A G C A G G T G C T C T C

Cons

- Combinatorial
- only get proteins not genomes or relationships between genomes

Steinegger et al. *Nat Methods* **16,** 603–606 (2019). https://doi.org/10.1038/s41592-019-0437-4

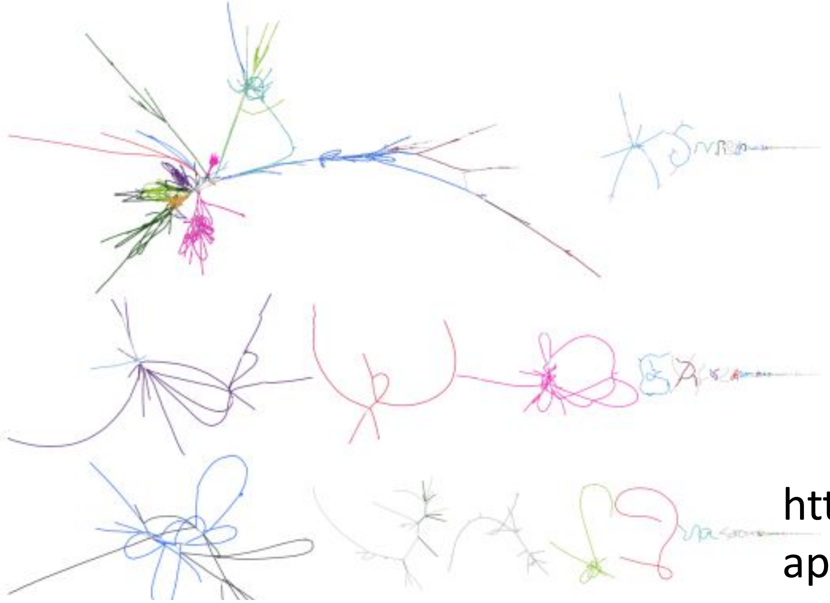
Metagenome assembly graph analysis

Pros

- Contains all the reads
- Mostly organized similarly to how those sequences occur in a genome

Cons

- Messy
- Big
- Tool space is still developing



https://tylerbarnum.com/2018/02/26/how-to-use-assembly-graphs-with-metagenomic-datasets/